关于新华社数据交换平台发展的思考

摘　要：新华社数据交换平台作为新华社的一个重要技术系统，已经平稳运行超过 20 年。近年来，随着新媒体技术及移动互联网的发展，新闻信息的生产及传播方式都发生了较大变革，既有的系统架构及业务模式越发难以满足新增的业务需求。现通过对数据交换平台现状的梳理分析，找到痛点，提出日后发展改进的方向和策略，顺应技术潮流，更好地为新闻信息传播事业提供服务。

关键词：数据传输；数据服务；统一存储；分布式部署
中图分类号：TP39　　文献标识码：A
文章编号：1671-0134（2019）06-099-02　　DOI：10.19483/j.cnki.11-4653/n.2019.06.029

文 / 张驰

背景

目前，新华社总社范围内的技术系统多达数十个，主要包括采编系统、发布系统、OA、数据库及新华网等，此外还有 31 家国内分社、11 家海外总（大）分社。上述系统组成了一个以新华社总社为核心，规模庞大的分级式业务网络。随着新华社全媒体新闻事业的蓬勃发展，相关技术系统的数量在增加，随之而来的各技术系统内部、社内各技术系统间、新华社技术系统与外部技术系统间的信息流转越发频繁，不同网域、不同系统、不同格式的信息共享、交换需求日益增多。

新华社通信系统始建于 20 世纪 90 年代，20 多年来一直作为新华社的核心技术系统之一，主要承载着总社各系统、总社内外网之间、总社与国内外分社之间、与社外系统之间的数据交换工作。系统内部处理的业务包括新华社文字、图片、音频、视频、多媒体等成品数据，以及外媒新闻、外部接入的异构数据等。多年来，随着业务发展，通信系统也在持续进行不同程度的业务扩展及迭代，逐渐演变为一个覆盖面广、实用性强的数据交换平台，为全社乃至相关社外机提供基础数据传输服务。

1. 现状及痛点

2010 年前后，世界大步迈入移动互联网时代。新闻生产及传播的业态也发生了巨大变革，与之相关的技术系统必须快速响应，顺应潮流。

对此，数据交换平台作为新华社的基础服务提供者，势必需要做出调整，找到制约自身发展转型的短板，对症下药。

1.1 系统架构繁冗

近十数年来，随着新华社新闻事业的快速发展，为了对接采编部门及终端用户需求，先后涌现出不少技术系统。这些系统的网络架构各不相同，同时，各系统间均存在个性化的数据交互需求。基于此，数据交换平台通过部署在新华社内网、DMZ 区、外网、绿区交互区、绿区应用区及私网等六个网络区域的节点机（每个网域均部署有一到多台节点服务器）完成本网域内、跨网域间的数据汇聚、格式转换、数据分发等数据交换工作。目前，数据交换平台中担负数据交换业务的节点服务器多达 20 余台，各网域的接入交换机 10 余台。硬件数量多，服务器主备机之间采用一对一冷备方式，业务布局分散，给系统管理员日常运维造成了不小的压力。

1.2 业务模式相对单一

多年来，数据交换平台所提供的数据传输及数据处理等服务，无论是在内部技术系统之间还是与外部用户之间，基本均围绕“文件”这一种数据形式展开。但随着移动互联网技术的蓬勃发展，新闻信息的传播方式也相应发生了巨大改变。比如我们看到通过消息驱动、借由 API 接口进行数据交互的技术路线越来越多的出现在各类应用场景中；RSS，数据订阅等数据获取及发布模式也被广泛采用。相较之下，数据交换平台沿用多年的仅基于文件及目录的数据传输模式已无法很好地满足业务需求，制约了自身的发展。

1.3 应用程序功能的健全性及规范化

数据交换平台作为传输中枢，上下游间交互的技术系统数量繁多，各系统在数据传输的过程中或多或少都存在一些个性化的需求，如所采用的传输方式不同（socket 或 FTP），所采用的操作系统类型不统一（windows，linux，solaris），文件落盘的方式要求不尽相同（是否按日期结构落盘，是否按照语种落盘，是否落多个实体等），甚至当涉及国际网域间传输时的网络条件是否要考虑数据校验及断点续传等。为了满足不同的技术需求，提供个性化的服务，数据交换平台内的数据传输处理程序先后衍生出不同的版本，各版本在主要功能上类似，但细节上均有差异，不易于维护，在后续业务部署时容易造成混乱。

1.4 缺乏统一高效的业务监控及管理手段

如前文所述，数据交换平台目前所辖主要传输节点服务器逾 20 台；平台内大部分应用程序均基于 C 语言编写，同时搭配一些 shell 脚本。基于这些原因，当遇到日常系统故障排查及业务调整，需要系统管理员根据业务资料在数据链条中涉及的每台服务器上通过命令性的方式进行操作，效率较低且容易出错。

2. 未来调整的方向

以面向服务体系结构 (SOA) 为框架，采取松散藕合方式构建，提供数据接入、格式转换、传输、回传、查询、检索等不同的服务；能够提供跨平台数据交换服务，

能够对数据接入、转换和传输过程实现集中统一控制和规范管理；针对每一条数据从接入系统开始，进行全流程的管理和配置。

2.1 系统架构设计

数据层面引入统一存储。当前，数据交换平台系统架构庞杂的一个重要原因在于被传输的数据均存放于各系统的本地文件系统中，因此需要在各网域部署传输节点，将同一份数据在不同网域间往复传输。统一存储（如NAS）的引入，可以为此类问题提供一个解决方案。存储网络作为区别于服务器业务网络独立存在的一张网，可以满足位于不同网域的服务器同时接入同一个存储网络，实现数据共享，在提高数据访问时效性的同时大幅减少数据在服务器业务网间传输的需求，节省网络资源。此外，NAS 本身自带访问权限控制功能，通过对不同的接入用户的读、写、执行权限进行细粒度的配置，可以确保基于统一存储上的数据安全性。因此，仅需要为暂时无法接入统一存储的网域部署节点机即可，服务器的部署数量上与之前相比可大为减少。

计算资源、服务层面采用分布式部署，集群模式。依托统一存储，无论稿件数据还是系统应用数据均可以方便地在服务器之间实现共享。因此，数据交换平台的计算资源完全可以按照服务功能进行分布式部署，以集群的方式实现。这样做的好处在于：首先，按照不同的服务功能进行分布式部署，可以使不同的应用模块间的耦合度相对松散，在对业务进行管理时逻辑更加清晰，快速定位问题所在；其次，由于实现了数据库共享、配置文件共享，服务器层面可以很容易做到“双活”乃至“多活”，相比于之前传统的服务器一对一冷备，这种集群工作模式使业务运行的稳定性显著提升，一旦一台服务器出现应用故障甚至宕机，集群中的其他服务器可以立即完成接管，业务完全不受影响，保证延续性。此外，集群模式为实现业务负载均衡提供了基础，这对于一些流量集中的核心业务节点来说是十分重要的。

2.2 服务模式的升级

在过去以“文件”为中心的业务模式基础上，增加并重点发展以“消息”为核心的业务模式。依托成熟的消息中间件，数据交换平台内部各应用之间、数据交换平台与外部系统之间的数据交互和服务调度都可以通过消息来实现。前文提到的“分布式部署”“服务器集群”就是通过消息驱动业务最直观的实例。

将数据交换平台常用的功能模块，如格式转换、数据分发甚至数据传输等，封装成服务，通过发布的 API 接口供各相关系统调用。从关联系统的角度看，通过调用数据交换平台的服务接口拿数据，在拿到数据的同时也可以根据自身需求开发或部署相关的应用对数据进行灵活处理；对数据交换平台来说，仅需要维护平台内的基础功能模块并确保接口的稳定即可，不需过多考虑关联系统的个性化需求。这样使得系统间的边界更加清晰明确。

2.3 系统应用的健壮性和稳定性

将程序进行重构，基于 java 和标准的 J2EE 规范实现，能够保证应用跨平台平滑部署和实施，不再受操作系统平台的局限；同时，在对有关数据传输程序的重构过程中，将个性化的功能通过丰富配置文件内容项进行设置，主程序中对应预留好相关功能入口即可。这样可以基本确保系统管理员在对业务调整时不需要对主程序进行太多修改，只需要重点对配置文件进行操作即可。这样可以保证应用程序功能及版本的相对稳定统一，同时也易于将应用模块打包，或以 agent 的方式部署在相关系统的接口机上。

2.4 管理监控功能的升级

接入 ELK 实时日志分析查询平台，可以使日常业务监控更便捷高效。

ELK 是三个开源软件的缩写，分别表示 Elasticsearch、Logstash、Kibana，它们都是开源软件。新增了一个 FileBeat，它是一个轻量级的日志收集处理工具，以 Agent 的方式装在需要收集日志信息的服务器上，在各个服务器上搜集日志后传输给 Logstash。

所有的日志数据采集并存储后，Kibana 可以为 Logstash 和 ElasticSearch 提供日志分析友好的 Web 界面，可以帮助汇总、分析和搜索重要数据日志。

为了让接入 ELK 日志平台的数据使用起来更加高效，查询及定位问题更加准确，系统内各应用的日志输出均必须遵循统一的标准。

系统硬件监控：对系统内所辖服务器的硬件情况进行监控，主要包括但不限于硬盘使用空间、内存使用率等。这部分信息都可以通过提取操作系统的 message 信息及执行简单的 shell 命令获得，并生成日志文件。

业务监控：在重构系统内部各模块的程序时，要按照统一的格式标准输出日志。通过对日志内与业务故障相关的字段进行直、简洁的设定，以求在接入 ELK 平台后，能够精确快速地检索出故障信息。由于每台服务器的日志信息都汇集到一起，因此，在日志平台查询时能够做到集中展示，甚至通过一条数据在不同服务器上的日志留痕，将业务链条串起来，帮助系统管理员快速定位问题所在，并及时进行处理。媒

参考文献

[1][加]托马斯.埃尔，李东，李多译，SOA 架构：服务和微服务分析及设计[M].北京：机械工业出版社，2017,（2）：10–26，40–72。

[2][美]埃克尔，侯捷译，JAVA 编程思想[M].北京：机械工业出版社，2002（2）：651–667.

[3]倪炜.分布式消息中间件实践[M].北京：电子工业出版社，2018（1）：23–42.

[4]冰心无影.ELK 日志分析系统详解[EB/OL].2017–04–24，https：//blog.51cto.com/12854546/1918773.

（作者单位：新华社技术局）